Homework I

Due date: 11:59pm, Feb. 11, 2024

For questions, please contact Cole Guo at zhishguo@tamu.edu

Note: For written questions, you can either turn in a scanned copy of your handwritten answers or a PDF file of your answers. For programming questions, you need to submit your code. Please put your code and written answers in a zip file and submit it on Canvas. Name the submission package as hw1_Lastname_Firstname.zip.

Problem 1: Probability (15 Points)

Suppose that we have three coloured boxes r (red), b (blue), and g (green). Box r contains 3 apples, 4 oranges, and 3 limes. Box b contains 1 apple, 1 orange, and 0 limes. Box g contains 3 apples, 3 oranges, and 4 limes. If a box is chosen at random with probability $p(r) = 0.2$, $p(b) = 0.2$, $p(g) = 0.6$, and a piece of fruit is removed from the box (with equal probability of selecting any of the item in the box), then what is the probability of selecting an apple? If we observe that the selected fruit is in fact an orange, what is the probability that it came from the green box?

Problem 2: Data likelihood (20 Points)

Assume we flipping a coin with probability of heads to be p and tails $1 - p$. If we flip it 5 times, what is the probability P that we observe a sequence of $\{head, tail, head, tail, head\}$? What is the p that maximizes this probability? (hint: since $f(x) = \log(x)$ is an monotonically increasing function, whatever p value that maximizes $\log(P)$ also maximizes P.)

Problem 3: Matrix Derivative (20 Points)

Suppose $\mathbf{x}, \mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ to are vectors and $\mathbf{X}, \mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$ are matrices. Derive the following first order derivatives. Denote $\mathbf{X}_{i,j} \in \mathbb{R}$ to be the element at i-th row and j-th column of $\mathbf{X}$.

1) $\frac{\partial \mathbf{x}^\top \mathbf{a}}{\partial \mathbf{x}}$
2) $\frac{\partial \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{x}}$
3) $\frac{\partial \mathbf{a}^\top \mathbf{X}\mathbf{b}}{\partial \mathbf{X}}$
4) $\frac{\partial}{\partial \mathbf{X}} \|\mathbf{X}\|_F^2$, where $\|\mathbf{X}\|_F = \sqrt{\sum_{i=1}^{N} \sum_{j=1}^{N} (\mathbf{X}_{i,j})^2}$ is the Frobenius norm of $\mathbf{X}$.

Problem 4: Matrix Rank and Inverse (15 Points)

Let $\mathbf{X}$ denote a matrix:

$$\mathbf{X} = \begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix}. \tag{1}$$

What is the rank of $\mathbf{X}$? Is it invertible and why? If yes, please derive its inverse.

Problem 5: Python Programming: Finding the Nearest Neighbor (30 Points)

Dataset In this problem, you will work on a hand written text recognition task on the USPS data preprocessed by LibSVM[1]. The data is included in the homework package (training data in "usps" and testing data in "usps.t"). In the dataset, the first column is the label, and the remaining columns are features in the form of (feature_index:feature_value). For a detailed description of the features and task, you can refer to the original paper[2]. You can load the data by sklearn[3].

Environment Setup If your laptop already has a Python environment, you can install sklearn using command "pip install scikit-learn". Otherwise, you can install Anaconda or Miniconda environment[4]. Then you should be able to access the Anaconda either through the installed Navigator or command line [5]. Then you can create/manage your own environment as [6].

Problem Background For any two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, let $D(\mathbf{a}, \mathbf{b})$ denote a distance metric that measures the distance between the two vecots. Some commonly used metrics include:
i) For two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, Euclidean Distance is defined as

$$D(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{i=1}^{d} |\mathbf{a}_i - \mathbf{b}_i|^2}, \tag{2}$$

ii) Manhattan Distance is defined as

$$D(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{d} |\mathbf{a}_i - \mathbf{b}_i|, \tag{3}$$

and iii) Cosine distance is defined as

$$D(\mathbf{a}, \mathbf{b}) = 1 - \frac{\sum_{i=1}^{d} \mathbf{a}_i \mathbf{b}_i}{\sqrt{\sum_{i=1}^{d} \mathbf{a}_i^2} \sqrt{\sum_{i=1}^{d} \mathbf{b}_i^2}}. \tag{4}$$

[1] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html#usps
[2] J. J. Hull. A database for handwritten text recognition research.
[3] https://scikit-learn.org/stable/datasets/loading_other_datasets.html
[4] https://docs.anaconda.com/free/anaconda/install/index.html
[5] https://docs.anaconda.com/free/anaconda/getting-started/
[6] https://conda.io/projects/conda/en/latest/user-guide/tasks/manage-environments.html

Requirements 1) Denote a data point by $(\mathbf{x}, y)$, where $\mathbf{x} \in \mathbb{R}^d$ is the feature vector, and y is the label. For the first 10 data points $\mathbf{x}_i$ in testing set, you are asked to find its nearest neighbor $\mathbf{x}_j$ from the training set according to the distance metrics defined above, i.e., $\mathbf{x}_j$ is the feature in the training set that minimizes the distance to $\mathbf{x}_i$ (Note: do NOT involve label y in the calculation of distance).

2) For all the $(\mathbf{x}_i, y_i)$ to $(\mathbf{x}_j, y_j)$ nearest neighbor data pairs that you identify, how many of them (in percentage) have the same labels, i.e, $y_i = y_j$?

3) You need to implement all three distance metrics mentioned above separately, i.e, you need to find nearest neighbor under each metric. Therefore, in question 2) you need to report the percentage under all three metrics, separately.

4) Include your code in the submission package, but do NOT include the dataset.

Problem 6: Probability Theory (Optional, 20 Bonus Points)

There is a rare disease that only happens to 1 out of 100,000 people. A test shows positive 99% of times when applied to an ill patient and, 1% of times when applied to a healthy patient. Please answer the following questions.

1. What is the probability for a patient to have the disease given that the test result is positive?
2. What is the probability for a patient to have the disease when he did two tests and both of them show positive? Assume that two tests are conducted independently.
3. Assume that the patient keeps on trying the test, what is the minimum number of tests that the patient has to try to be 99% percent sure that he is actually ill? Assume that all tests are conducted independently.